

A U.S. Decennial Census Browser Shiny application: Visualizing the African American population over time

Kiegan Rice, Iowa State University

Introduction

Census data provide an important snapshot of information about a country at different times throughout its history, and their value is difficult to overstate. While history books present a narrative of the events that occurred, students themselves often don't get to interact with the raw data in that learning environment. Clean and accessible census data allow the exploration of different demographic groups over time, or investigations of a particular period of time and what the demographic and economic landscape looked like in the past. Even today, data about the world around us opens a pathway for learning more about places we haven't been and groups of people with whom we may not usually engage.

From an early point in the United States' history, there were many "eminent men of science" who recognized the value of the census data and worked to aggregate and present those data. Francis A. Walker's "Statistical Atlas of the United States", based on the 1870 census, was an impressive effort in aggregating population data to present them in a visually appealing way [3]. Although the Census Bureau's "Statistical Atlases" eventually stopped being created, they were an important start to the effort of visually presenting census data to a wider public (see Figure 1).

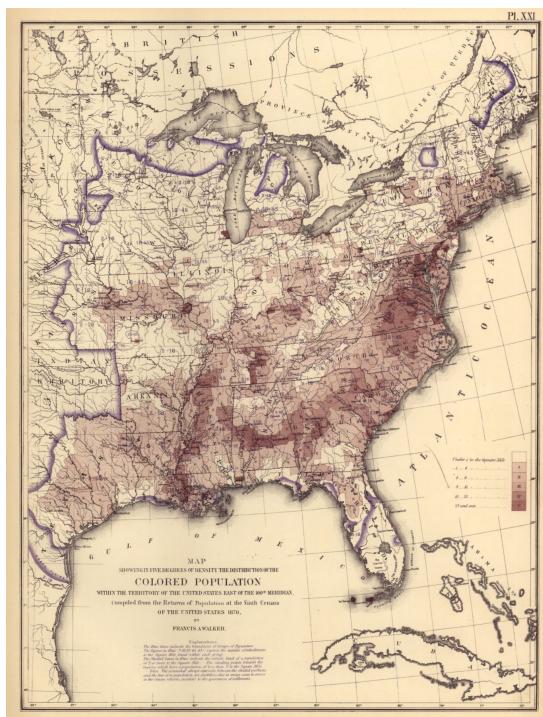


Figure 1: A visualization of the African American population in the United States from the "Statistical Atlas of the United States".

Today, as methods of data analysis and visualization continue to be developed and improved, access to census data allows us to look back on that history and explore,

synthesize, and visualize the information. When aggregated and presented in a clear manner, viewers can learn more about patterns in many different parts of the population. A 2007 paper about the Statistical Atlas discusses just how much information is present in those original charts and demonstrates several other ways of presenting the information that can be found therein [2].

Ever-improving visualization and data-wrangling methods in R [4] give those interested in statistical graphics a wealth of opportunities to explore and learn from data; in particular, incorporating user interactivity using Shiny has revolutionized the way statisticians share and communicate information [1]. However, it is difficult to make use of these tools with census data if those data are not available and easy to explore in one location.

An inherent problem in census data is that a country's census changes over time; the variables collected, how they are collected, and even the locations on which they are collected are updated as the country is formed, and subsequently grows and changes. The United States census data are no exception to this rule. In little under two and a half centuries, the census has taken many different forms. Data on occupations have transformed as the employment landscape has changed; new states have been formed, the most recent being within the last 100 years; definitions of various demographic groups and the terminology used to describe them have been updated as the demographic makeup of the country has changed. Each decennial census brings a different set of variables to the table. Sometimes these variables are new things the Census Bureau is interested in learning about, while sometimes they remove variables that are no longer relevant or whose information is captured elsewhere.

Unfortunately, because the founders of the U.S. Census were unable to see 200 years into the future, those interested in working with census data are left with quite the inescapable mess. If you want to focus in on a particular demographic group and their journey as part of the population of the United States, you may have ten or more different variable names to describe that one group over the course of the census from 1790 to 1960 - and that is just for a single group! Of course, we cannot just simply change variable names to match our own research needs. It is important to keep the data in their true form and be respectful of the way in which the population was defined at different times throughout history, even if our instinct may be to 'clean' the data by changing variable names.

This, of course, leaves the user with a wide variety of variables that are far from consistent across years. In order to track one demographic group across years - let alone many groups - a clean user interface that helps users see exactly what information is available to them is a necessity. To streamline the process and assist researchers in finding out to what information they have access and what information they lack, a U.S. Decennial Census Browser is

presented as a Shiny application, with the ability to explore available variables within a certain year or across several years. Additionally, a Visualization tab will allow users to immediately visualize the variables they have picked across different years to get an initial look before downloading the data in the form of ‘tidy’ comma-separated values (csv) files.

Data Access

Should I include a small section about Data Access? Or should I condense the introduction and add where the data is coming from in that initial section?

The Census Browser Shiny application

The Shiny application has three main points of focus, split into separate tabs. The first tab, ‘Single Year’, allows users to explore all available variable names for a single year of the decennial census at a time. After choosing a particular year, users can look for up to two specific variables of interest by using two available text search bars. This will narrow down the possible variable names, allowing for a more narrow focus in a given year.

The second tab, for ‘Multiple Years’, brings in the ability to compare availability of variables across several years. Users may select a range of years and, if they so choose, narrow down the given results with a text search term. These results are presented as a table with all chosen years as ‘columns’, while all relevant variable names appear as ‘rows’ in the table. An ‘X’ indicates the presence of that particular variable name in the corresponding year. A totals column to the far right allows users to not only see in how many years each variable appears, but also order the relevant variables by that frequency.

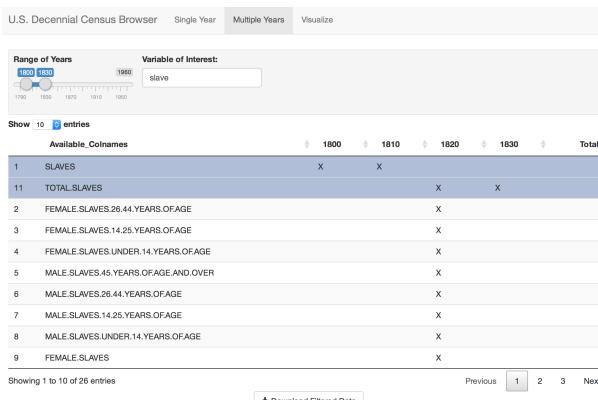


Figure 2: Resulting view of the ‘Multiple Years’ tab after restricting the range of years to between 1800 and 1830, and searching for the term ‘slave’.

The final tab, ‘Visualize’, creates an on-the-spot visual representation of the data selected in the ‘Multiple Years’ portion of the application. Users may choose to simply separate data by year, and sum over all variables chosen within each year, or facet by both year and variable chosen.

Depending on the range of years selected by the user and the number of different variables chosen, the latter can easily become a very large grid environment. However, it still allows users to see several things about the particular variables and years they are looking at (see Figure 3).

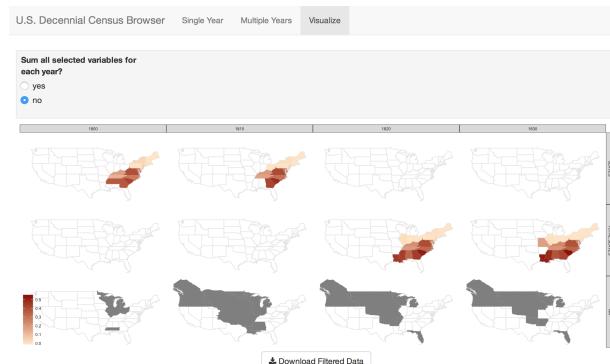


Figure 3: A visual representation of available variables for total number of slaves between the years 1790 and 1840, with period-accurate mapping for state boundaries.

First of all, every plot has the outlines of the 48 contiguous states as of 1999. Above that base, the outlines for each state are plotted as they were in the year represented. This means that users can see varying state boundaries across years - however, since these are the areas data were actually collected on at the time, they are the true representation of the geographic regions for the demographic groups chosen. In some years, there are states that still hold Territory status, or are not yet included in the decennial census. These will show up in a bottom row, since they have NA values for the variables. However, it is still useful to see where those boundaries existed and see how they change over time as you pan across the columns.

Secondly, users can see how variable names change over time. They can see this during the initial selection of variables in the ‘Multiple Years’ tab; however, now they can see differing variable names for an equivalent or similar demographic group and still directly compare the values on the maps.

A Closer Look at the Data

After ordering by how many years each variable appears in, the available variable names show a serious lack of continuity. The four ever-present variables (**Year**, **State**, **TOTAL.POPULATION**, and **Type**) each appear ten times, once in each year selected. After that, the next most common variable only appears in four of the ten years we have chosen. This highlights some of the hurdles that we will encounter when tracking any demographic group over time.

Since the focus will be on African Americans throughout U.S. history, the first search term needed is **SLAVE** in the early years of the U.S. census. The vast majority of African Americans were slaves when the United States was founded, so variables that count the number of slaves tell the story about the African American population in early U.S. history. Note that the ‘**SLAVE**’ categorization is the only term the U.S. census had related to African

Americans in the early years of the census, so it is also the only source of information available on that group for that time period.

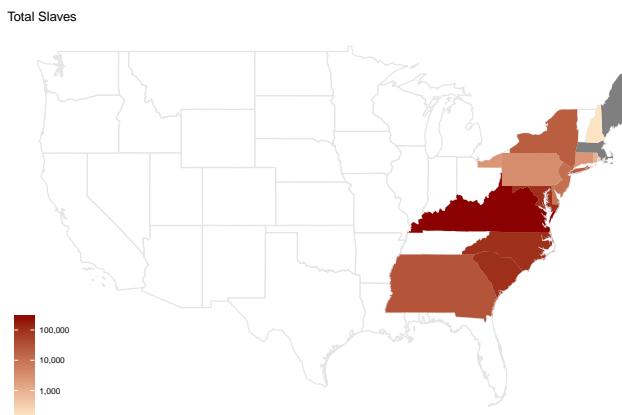


Figure 4: Total number of slaves per state in 1790, plotted on a continuous log scale. State boundaries for July 4, 1790 were gathered from ‘USAboundaries’ package.

With a visualization of the slave population in the United States in 1790 in hand (see ??), it is possible to look at subsequent years of the decennial census in a similar manner to see how the population changes over time.

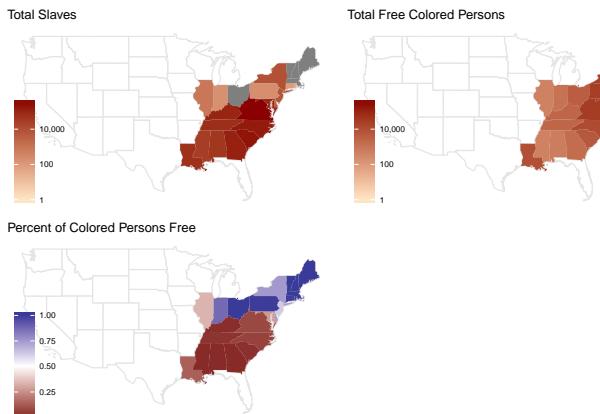


Figure 5: Total number of slaves, total number of free colored persons, and percentage of colored persons free in each state in 1820 United States.

Jumping to the year 1820, and again filtering on that particular year, we learn new information about the African American population. The census at this time includes a count of TOTAL.FREE.COLORED.PERSONS, as there are citizens who are African American or other minorities that are free, and not slaves. At this time, many minorities were grouped together in the “colored persons” category, and as before, we do not have any other record for African Americans except this categorization and the “slaves” categorization. However, since the 1820 census included records for both TOTAL.FREE.COLORED.PERSONS and TOTAL.SLAVES, we can also now visualize the percentage of African Americans in each state that were categorized as “free persons” as opposed to slaves (see ??). Knowing that both of these variables are available allows for a powerful comparison by visualizing the growing divide between northern and

southern states in the United States at that time.

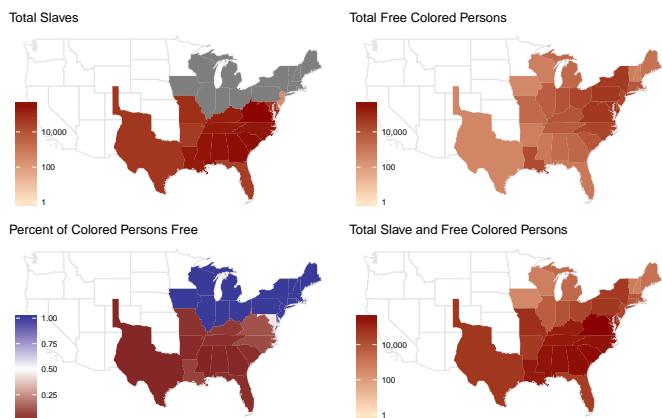


Figure 6: Total number of slaves, total number of free colored persons, and percentage of colored persons free in each state in 1850 United States.

Moving forward in time, even more information can be gleaned about this population. When we subset on the year 1850, we see that although there is still a TOTAL.SLAVES column in the record, many states no longer record this variable; by then, they had already abolished slavery. We can transform these NA values into zeros to account for this difference in the data and allow us to still calculate the percentages of free African Americans. Starting in 1850, we can also start investigating what the total African American population looks like in each state, beyond just the size of the slave population and freed population (see ??).

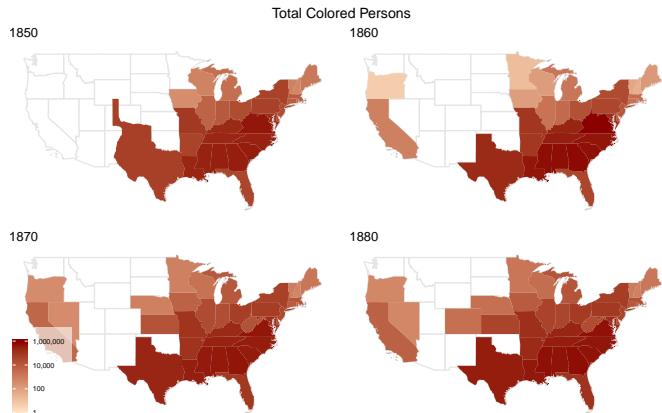


Figure 7: Total number of colored persons per state in 19th Century United States.

The total African American population is particularly interesting moving forward after slavery is abolished in the 1860s and some migration begins to occur out of certain areas. The balance of the total African American population in each state begins to change as the United States transitions into the post-slavery period (see ??). We also see how the boundaries of states change and new states are formed, with the United States gaining new states at each census.

A period of growth and change can be seen at the turn of the 20th century as well, with a dramatic shift in the

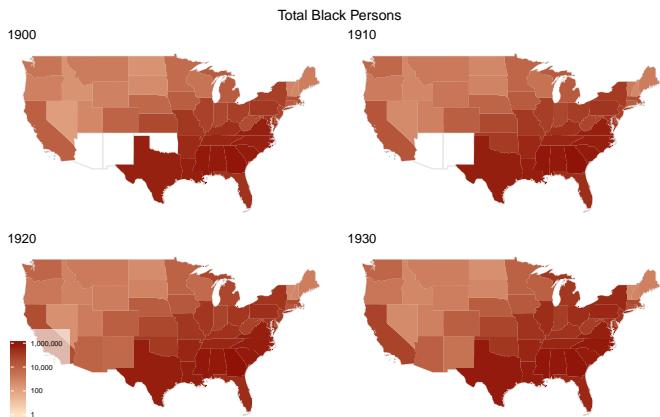


Figure 8: Total number of black persons per state in 20th Century United States.

African American population out west to California and with the balance in the east and midwest shifting slightly further north (see ??). Although the highest density of colored persons is still in the southern states, we can see that of a northern migration is taking place.

Conclusion

The decennial census is a rich source of historical information. We have built a tool that can help extract and visualize that information to, for example, understand how demographic attributes of the U.S. population have changed over time.

Although there is no county-level information available and the data for the census browser go only through 1960, users are able to efficiently explore variables of interest at different time points in U.S. history and tell a visual story about a growing nation and a changing population. This connectivity across years and ability to assess what information was available when, all in one browser, have the potential to save time and effort of researchers interested in delving into United States history.

The structure of our browsers also helps determine not only the variables that are available in a given year, but also where information is missing for some states on a particular variable. Finding these differences is easier when all relevant variables are combined in one data frame, which can be summarized and searched all at one time.

References

- [1] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPher-
son. shiny: Web application framework for r., 2017.
- [2] H. Hofmann. Interview with a centennial chart.
CHANCE, 20:2:26–35, 2007.
- [3] U. S. C. Office and F. A. Walker. *Statistical atlas
of the United States based on the results of the ninth
census 1870 with contributions from many eminent men
of science and several departments of the government.*
1874.

- [4] R Core Team. R: A language and environment for statistical computing. 2015.

Further Reading